**Johns Hopking Genomics FACILITIES and RESOURCES**

## A. Laboratory:

**1812 Ashland:** In preparation for and in response to the ever-expanding variety of genomic services required to support human genetics research, we established Johns Hopkins Genomics (JHG, jhgenomics.jhmi.edu), a new initiative of the JHU School of Medicine to combine its research and clinical genomics expertise, and is a joint effort between the McKusick-Nathans Institute of Genetic Medicine (IGM) and the Department of Pathology. The formation of JHG is a multi-million dollar investment for the School of Medicine. A core group of more than 100 faculty and staff occupy the 2nd floor (25,000 square feet) of the newly constructed 1812 Ashland building in the East Baltimore Science+Technology Park adjacent to the Johns Hopkins medical school campus. Wet laboratories occupy ~60% of the new floor and an open office suite ~40% of the total square footage.

This new arrangement is a true melding of groups; all laboratory space is shared and CLIA compliant, and the open office space configuration has fostered collaboration among the clinical, research, informatics, and analytic faculty and staff (see Appendix 8 for space diagram). JHG consists of the staff and expertise of two major research labs – CIDR and the Genetic Resources Core Facility (GRCF) – and three clinical genomics labs: the DNA Diagnostic Laboratory (DDL) for inherited germline diseases (DDL, in operation since 1979), the Pathology Molecular Diagnostic Laboratory (MDL) for cancer genomics (in operation since 1991), and a new Clinical Genomics Center (CGC) to offer whole exome and whole genome clinical tests. The CIDR/GRCF infrastructure allows for the processing of large sample numbers and the equipment and informatics expertise required for whole exome sequencing (WES) and whole genome sequencing (WGS) services. In total, the combined laboratories have over 150 faculty and staff providing a variety of services, the most relevant of which are CLIA/CAP sequencing and genotyping for inherited and somatic mutations. Both existing diagnostic laboratories have a rich history of supporting clinical research studies and, by combining with CIDR and the GRCF, will be enabled to support much larger scale clinical translational research studies as required. The DDL is a core lab (Genetics Translational Technology Program) of the Johns Hopkins Institute of Clinical Translational Research and performs clinical-grade testing for approved translational research projects. The MDL has close ties to Kimmel Cancer Center researchers and has supported over 65 clinical translational research studies.

The 1812 Ashland laboratory space is approved for BioSafety Level 2 work and supports specimen intake and quality control; DNA and RNA isolations; tissue dissections; Illumina Infinium, Affymetrix Axiom, Fluidigm, Sequenom-based research and clinical genotyping; Sanger sequencing; and Ion PGM and Proton, Illumina Miseq, Hiseq2500, and NovaSeq6000 research and clinical next-generation sequencing. In addition, long read sequencing is available on the Oxford Nanopore GridION platform. All laboratory space is CLIA-compliant, physically separated, and organized into color-coded levels, stratified down to the level needed to segregate the sequential common primer PCR steps during the sequencing library prep workflow. The blue labs are all pre-PCR/pre-amplification lab space for both genotyping and sequencing. The green labs are post-PCR/post- amplification/PCR (1) lab space for genotyping and sequencing. The orange lab is for the PCR (2) of library capture. The red labs are for post-PCR steps of sequencing. The MDL, DDL, and CGC are all CAP- and CLIA-certified laboratories with a tremendous variety of clinical and research assays in use or being developed. Combined, they currently perform over 10,000 clinical tests per year and will be expanding their clinical exome, whole genome, and large panel-based NGS services through the efforts of the new CGC. Please see the JHG Equipment list for a complete list of available resources.

**Other laboratories:** The GRCF units that are not located at 1812 Ashland occupy ~4,000 square feet of laboratory and office space on the 10th floor of the Blalock building, ~6,000 square feet at an offsite location (Lighthouse Point), and ~2,000 square feet of LN2 repository space at the Fayette St. loading dock. The JHU Cell Center tissue culture laboratory is a CAP-accredited facility on the Johns Hopkins Hospital campus. The facility includes 3 biological safety cabinets and 6 ThermoFisher humidified incubators in 2 negative air pressure rooms. Cells are quantified and quality control monitored by way of 2 Vi-Cell XRs. The Rees Environmental Monitoring System continually monitors the cell culture conditions. Single cell genomics is accomplished using Fluidigm's C1 Single Cell Auto Prep and 10X Genomics Chromium platform. The JHU Biorepository is a CAP-accredited facility that houses 2 Taylor Warton LABS 80K LN2 vapor phase freezers, 6 Taylor Warton LABS 40K LN2 vapor phase freezers, 2 MVE High Efficiency LN2 vapor phase freezers and 2 ThermoFisher 20CUFT chest freezers. For cryogenic transport of biospecimens, the Biorepository utilizes 2 portable LabRep Co. Liquid Nitrogen Cryocarts. Freezer security and integrity are maintained through the Rees Environmental Monitoring System, with an automated Centron monitor. Repository inventories are maintained through the Freezerworks database system. The Fragment Analysis Facility (FAF) houses equipment for both STR and SNP genotyping, nucleic acid extraction, qPCR, and mycoplasma detection services. Equipment includes a Taqman7900HT, four Veriti (Applied Biosystems) thermocyclers, a Qiagen Autopure DNA extractor, a QIAcube that provides automated processing of Qiagen spin columns, a Hamilton Microlab STAR liquid handling workstation, Integra Biosciences VIAFLO 96, and a CheckScanner for reading mycoplasma MycoDtect™ microarrays. The lab also has -80 and -20 freezer storage, a variety of electrophoresis and centrifugation equipment, three Nanodrop spectrophotometers (two single, one multiple), a Spectramax Gemini XS UV plate reader, a FragmentAnalyzer (Advanced Analytical Technologies) for nucleic acid quality, quantity and sizing analysis, and a Millipore water purification system.
The JHU DNA Services laboratory within the GRCF is located in ~400 sf the 10th floor of the Blalock building on the Johns Hopkins Medical Campus. This laboratory houses equipment for qPCR, digital PCR and medium throughput genotyping (2 QuantStudio 12K

Flex machines, one Taqman 7900HT and one QuantStudio 3D), Sanger sequencing (2 3730XL machines), and Pyrosequencing (Pyromark Q24).
The Cytopathology Lab occupies ~3,500 square feet of wet-lab space in the Park Building. Facilities are available for cell culture, microscopy, classical cytogenetics, FISH, and DNA microarrays. GeneScan™ analysis software and other design and analysis applications are available on a secure, networked lab information system.

## B. Informatics:

The informatics resources of Johns Hopkins Genomics are primarily comprised of the robust capacity of the existing JH CIDR group, supplemented by the clinical and research bioinformatics resources of the MDL and the DDL. All of these resources are currently located in the 1812 building.

**B.1. CIDR Informatics:** At JHU CIDR has continuously expanded and improved informatics infrastructure, including data movement and storage, computational and network capacity, and server room space, cooling, and power. The dramatic increase in CIDR services and sample numbers over the last few years produced concomitant increases in data production rates and volume, requiring rapid expansion of computing and storage equipment. The server room in 1812 Ashland is electronically secured, with full environmental controls and fire suppression. Over 170kVA of UPS capacity provides continuously filtered power and, in the event of external power outages, several minutes of bridging power until the building generators activate.

The compute cluster consists of 36 fast compute nodes plus several many-core large-memory servers, totaling over 615 compute cores, 5TB RAM, and 52TB of local storage. It is administered via Bright Cluster Manager software, enabling superior management of computing jobs and resource distribution, along with bioinformatics software versioning for research use and version lockdown for the clinical informatics pipelines. A VMware virtualization platform consists of three Dell R730 servers totaling 72 cores and 192GB RAM, running nearly 100 virtual machines. Storage systems include a 65TB X-IO fibrechannel SANS for provisioning disk to servers, plus over 1.5 petabytes of network-attached storage (NAS) including 17TB of solid-state disk for metadata acceleration, in separate clusters for production data and archival storage. Included among these totals is a separate CAP-certified NGS bioinformatics analysis platform for the DDL, currently consisting of four Dell C6220 compute nodes (64 cores) and four 72NL storage nodes (260TB disk). A Quantum i500 tape platform running CommVault Sympana backup software, consisting of ten ultra-fast LTO-5 tape drives and 211 tape slots in a robotic library, serves all these systems.

The local network is built around six new Cisco 10-gigabit optical core switches (140 ports total) and 13 1-gigabit copper edge switches, providing massive data movement capability protected by a Palo Alto Enterprise controlled Firewall. Our primary Internet connection is via multiple gigabit connections to the JHU network supplemented by a 10-gigabit connection (40Gb uplink) to the JHU research network for institutional data movement; JHU also has a 10-gigabit connection to Internet2. Currently, data distribution is via secure connections from an Aspera cluster at up to 300 Mbps per session.

The JHU research network also provides high speed access to the new Maryland Advanced Research Computing Center (MARCC) on the Hopkins Bayview campus. This facility offers a massive compute cluster of 19,000 cores, including 48 GPU nodes, 20 petabytes storage with 2PB in a fast Lustre parallel filesystem. MARCC is readily accessible for our use should the need arise for processing at a scale beyond what our in-house infrastructure can accommodate, and/or for pilot projects so as not to impact production infrastructure.

**B.2. DDL Informatics:** DDL operates on a secure LAN provided and maintained by Johns Hopkins Enterprise IT (IT@JH) Network Technology Services (NTS), with other informatics resources and services provided by the IGM IT team and by CIDR. NTS works closely with the IGM IT team for any network maintenance and performance enhancements. IT@JH also manages the Johns Hopkins 1830 Data Center, the physical facility that houses the lab's server hardware. The IT@JH Data Center managers also work with the IGM IT team to provide the best physical environment for all server equipment. The shared filesystem for DDL is a GPFS cluster filesystem optimized for high availability and designed to eliminate single points of failure. The GPFS cluster runs on Linux and serves DDL data to clients via Samba Windows File Sharing. Users must authenticate using a valid IGM user account in order to access the DDL Windows share; transactions (including file access and login attempts) are centrally logged.

Access to the DDL shared filesystem requires a valid IGM user account; authentication against the IGM directory server is encrypted via SSL. Passwords are stored on the directory server using a one-way hash. Data on the DDL shared filesystem is backed up nightly to a backup server using IBM Tivoli Storage Manager. Retention policy allows up to three prior versions of active files for retrieval; deleted data is retained 6 months. A second copy of all backup data is maintained on encrypted tape offsite for disaster recovery.

Client workstations are networked computers and do not store data locally. Client operating system and antivirus software receive regular security updates. All client machines require authentication against the JHU Active Directory service with a valid Johns Hopkins Enterprise Directory (JHED) account.

The IGM IT group currently provides the following services for the DDL :

- Client LAN Workstations (27)

- Shared Filesystem (GPFS cluster, accessible by lab personnel via PC or Mac clients)
- Sequence Pilot analysis software for Sanger confirmation sequencing
- High-Availability Web/Database Services for LIMS (with backup LIMS on separate hardware), billing, reporting

The JHG Informatics group currently provides the following services for the DDL:

- Aspera File Transfer Server
- Compute cluster for NGS analysis pipeline (BWA/GATK/Annovar)
- HP Z800 NGS data storage and analysis workstation (8 cores, 96GM RAM, 2TB disk)
- NAS cluster
- Windows Terminal Server

**B.3. MDL Informatics:** MDL's current informatics infrastructure is organized in two domains to accommodate the different requirements of clinical and research protocols, plus a high performance cluster (HPC) for immediate data processing.

- The Clinical domain consists of two Dell PowerEdge R510 storage servers with Intel Xeon processors and Intel 5500 chipset, configured with ZFS logical volume manager, with real-time backup using snapshot technology. These two servers accommodate MDL's Clinical data storage, one unit serving as primary onsite clinical data storage system and the other unit as secondary, fail-over storage system at an off-campus location. The two storage units are mirrored in real time and are configured and managed in a HIPAA-compliant manner.
- The Research domain consists of a 48TB Sun Fire X4500 storage server.
- Data processing for both the Clinical and Research domains is accomplished on a four-node HPC consisting of Dell PowerEdge 420s with Intel Xeon E5-2400 processors with PCIe 3.0 enabled expansion slots.
- The HPC is administered via Bright Cluster Manager software.
- The Clinical storage system and the HPC are connected via a 10gigabit Nexus 3064 TPORTS Ethernet switch with 32 ports connecting to the clinical network domain.
- MDL hosts virtualized Clinical and Research web servers for clinical sign-out and research collaboration purposes. All the storage for the virtual machines is maintained on the primary storage system.
- MDL uses customized SoftMolecular® LIMS (Soft Computer Corp.) for all its sample processing and QC tracking. This system connects externally with Department of Pathology information systems such as PDS and Care Fusion, as well as with Epic and other EMR systems.
- In addition to the above computing equipment, the MDL lab provides over 50 Windows PCs with twin Dell monitors, as well as Apple computers for the use of the workforce and laboratory equipment.
- Data emanating from non-NGS platforms is archived and managed on a Department of Pathology storage platform.

**B.4. Software Development:** Custom software development, which is key to the successful function and improvement of our complex, high-throughput informatics environment, is carried out by a team of five software developers, guided by a Sr. Bioinformatics Software Engineer, Team Lead with over 10 years of experience. The rest of the team is composed of a Software Engineer, a Sr. Programmer Analyst, and two Programmer Analysts, with over ten years of experience among them. In addition to software development experience, two of the software developers have previous experience working in genotyping and sequencing labs. This experience grants these developers exceptional understanding of both lab protocols and software usage, and of relevant biological processes. Large software projects are typically conducted by two or three developers, while small projects are pursued individually. That said, the team is highly collaborative, and cross-pollination of developers and ideas throughout the group is frequent. Even on small projects, developers benefit from sharing advice, ideas, and discussion.

Software developers at JHG employ an Agile methodology, which facilitates close collaboration with and rapid feedback from colleagues in other groups, particularly our bioinformaticians and statistical geneticists, who often create and refine early versions of programs and pipelines that are later re-engineered for production use. Because clients have access to software early in the development process, they are able to make feature requests often and to validate that the software provides the desired functionality. The adoption of an agile methodology has reaped numerous other advantages: during the requirements gathering process, small face-to-face meetings between developers and their collaborators allow for effective communication, and removes roadblocks caused by a lengthy and formal specifications gathering process.

JHG software projects are written using several technologies and languages. The most commonly used programming language is Java although there is production code written in Python, JavaScript, Perl, and Scala as well. Among the technologies used for various projects are JavaFX for GUI programming, Django and AngularJS for web development, Hibernate as an object-relational mapping (ORM) framework, OrientDB for graph-based databases, RMI for inter-JVM communication, Hadoop for big data analysis, and Accumulo as a secure NoSQL solution. Most software projects use MySQL as a database backend using a variety of intermediate layers, including JDBC, the aforementioned Hibernate, and a custom ORM written at CIDR. Some older LIMS systems still employ Oracle databases but those systems are currently being phased out in favor of MySQL.

All custom software is version-controlled using a local Git repository, which is backed up regularly. A few software projects have been transitioned to a remote Git repository using BitBucket. A remote repository has the advantage of providing a secure Git solution with lower local administrative costs. Storing code in a cloud-based repository has also allowed us to transition several projects to publicly available, open-source packages. Among these open source projects are CANNOTATE, SRA Submission, CIDRDB, CIDRUtils, and Calcordance. In addition, many JHG software projects have been presented at various meetings, including AGBT and ASHG. Please see the list of JHG Software Systems below for more information.

In addition to our major in-house software applications, there are dozens of smaller applications written by our developers. All software is developed using third-party libraries, such as those from the Apache Foundation, as well as with libraries of in-house developer tools. These developer tools include many GUI-based utilities, encryption and decryption tools, e-mail and communication utilities, and a framework for interacting with traditional SQL databases fluidly.

**B.5. Software Systems** (in alphabetical order)

**B.5.a JHG-Authored Software**

**ANNOVAR Reporting:** CIDRSeqSuite contains a series of utilities for performing annotations using ANNOVAR against a directory of VCF files and against any number of annotation databases (currently several dozen). These tools execute the annotations in parallel, then combine the data from the annotations and the VCF files into one large report, which can then be used to aid in variant interpretation.

**Bioinformatics Applications Manager**: staff typically access our information systems through Web or Java client-based graphical user interfaces created by the software development team. Most custom software is executed via the Bioinformatics Applications Manager (BAM). Written in-house using Java SE 7 and JavaFX, the BAM provides a single consistent desktop GUI for most applications, regardless of programming language or other constraints. The BAM features search functionality and a "favorites" area for each user to facilitate ease of use.

**Calcordance:** Calcordance (the name is a portmanteau of *calculation* and *concordance*) is a software application created to generate concordance information for the controls of a genotyping project as quickly as possible with a very large set of reference data. The tool uses custom compressed genotype files to quickly perform concordance between JHG data and data from the 1000 Genomes project. Calcordance can also perform comparisons to HapMap data when appropriate.

**CANNOTATE:** Now under development collaboratively with members of the Johns Hopkins Applied Physics Lab (JHU/APL), CANNOTATE is a software tool designed to manage and annotate genomic data RAVE (CANNOTATE is written in Java and uses Accumulo and Hadoop for data storage and processing. Accumulo, a NoSQL database, takes advantage of Hadoop's map-reduce framework to store and retrieve data across a distributed database. Accumulo also has the advantage of having cell-level security that is useful when data stored is of heterogeneous provenance. In Accumulo all data is stored as Key-Value pairs. Because both annotation and genomic variant data include positional information, the keys used for CANNOTATE data are based around this positional information. In initial tests on a modestly sized Hadoop cluster of four nodes, data annotation times ranged from 4 to 26 times faster than those of other annotation tools. Further development is currently supported by a JHU/APL internal R&D grant.

**CIDRSeqSuite:** Our CIDRSeqSuite sequencing data analysis pipeline was developed to automate a variety of next-generation sequencing analyses and is available for use by JHG. The current version of CIDRSeqSuite (v7) focuses on individual tasks and the interdependencies among them. When an analysis is submitted to the CIDRSeqSuite server, a record of all required tasks and their dependencies is stored in a relational database. Such tasks include alignment, variant annotation, and even the rapid parallel de-multiplexing of an individual tile from a flow cell lane and conversion to fastq. Tasks are submitted by the server process to an SGE-enabled cluster. As each task finishes, its status is stored in the database; any tasks whose dependencies are complete are then submitted. A retry policy resubmits failed tasks. Email notifications are sent upon the completion of analyses and when errors such as repeated task failures occur. Tasks can be submitted to the compute cluster via command-line tools or a graphical user interface.

**CIDRLIMS:** CIDRLIMS is a comprehensive LIMS framework meant to replace or augment all existing workflow-bases LIMS at CIDR and JHG, including those for sequencing and genotyping. CIDRLIMS is being developed in Java 8, with Gradle for module management and JavaFX for GUI front-end development. CIDRLIMS is designed for rapid development and deployment of new lab workflows with minimal lab disruption.

**CIDR WebLIMS:** To handle the workflow complexity and volume of data for large-scale GWAS genotyping projects, CIDR developed a system for sample tracking and quality control customized to our high-throughput environment and need for flexibility. Written in Java, JavaScript, and HTML with a MySQL back end, the WebLIMS tracks each sample through all stages of processing by recording all variables essential to quality control, with integrated data validation to reduce errors. The WebLIMS has been in daily production

use since 2006, with major changes in 2014. Data from the WebLIMS is presented to lab managers in various forms via several reporting tools. The WebLIMS also serves as a starting point for the genotyping data processing pipeline.

**Data Archiving Tools:** To facilitate management of the release of sequencing and genotyping data, developers created a Java desktop application to aid lab management in filtering, moving, copying, and deleting large quantities of files of diverse types. The application features a custom set of file manipulation modes that have been designed around the needs of sequencing lab managers. To match the dynamic requirements of lab data interpreters, the application's performance parameters are easily altered, with custom modes available to users.

**Genotyping Data Processing Pipeline:** Since 2006, genotype calling and QC have been performed by several iterations of AutoCall, a Java RMI and JDBC application integrated with the CIDR WebLIMS and Phoenix sample handling system. AutoCall is an automated analysis pipeline for data generated by Illumina's Infinium genotyping products. CIDR's pipeline augments the functionality of Illumina®'s "AutoConvert" utility, which converts scanner-produced IDAT files into GTC files. The GTC files are then read to calculate metrics including call rate, AA/AB/BB call frequencies, mean intensities for raw X and raw Y, estimated gender, and no-call and total call counts. Starting in 2015, the GTC files are also used to generate genotype files and PLINK files for project release via a set of Genotyping Data Release Tools mirroring some of the functionality of Illumina's GenomeStudio but designed to handle extremely large projects much more effectively than GenomeStudio, and in a fraction of the time. Our CIDR Autocall pipeline is currently being reworked to allow closer integration with the Genotyping Data Release Tools and to take better advantage of a new set of GTC parsing and analysis utilities.

**Miscellaneous:** Among the miscellaneous tools developed and used at JHG are flow cell and BeadChip inventory trackers, barcode printer management and alignment software, various administrative utilities, and a tool for interacting with delimited text files that resembles a simplified Microsoft Excel but has additional features.

**Phoenix:** Phoenix is a sample, problem and project tracking system developed in-house by the JHG software development team. Designed to accommodate the volume and complexity of sample management for SNP genotyping and next-generation sequencing work, Phoenix has been used in production since 2016. A web portal allowing investigators to communicate with Phoenix to upload files, check on the progress of their projects, and respond to problem reports has been developed and is actively used by NIH investigators. Phoenix provides the flexibility and sophisticated problem handling that enables us to deliver the high-throughput, high-quality service that makes JHG a unique national resource to the life science research community. Current work includes integrating project management features from the ProjMan system (see below) along with project release statistics and a publications database to track metadata about all completed projects. The Phoenix system also serves data to several other systems, including many QC report generating applications, LIMS systems, and the genotyping data generation pipeline.

**ProjMan:** This study and project management platform is an internally developed and maintained Java application built upon an Oracle database. It was developed around the concept of studies, projects (each data release is a project; each project is associated with a study, and there may be multiple projects per study), and principal investigators (each PI is associated with one or more studies, and therefore also with one or more projects). It provides a mechanism for creating, scheduling, tracking, and summarizing various aspects of projects. The programming team can add and modify tables, reports, and GUIs to fit the needs of the users. All ProjMan functionality, plus additional capabilities, have been folded into Phoenix and the data migrated to MySQL from its current Oracle back end.

**SRA Submission:** Many sequencing projects require submission to the Sequencing Read Archive at completion. This submission requires the creation of several comprehensive XML files prior to data transmission. Working directly with NCBI SRA staff, JHG developers have created an easy-to-use, wizard-based GUI tool for the creation of these files, which guides users through providing the required information via a simple drag-and-drop interface.

### B.5.b Integrated Third-Party Software Systems

**CIDR Troubleshooter:** The CIDR Troubleshooter is an electronic forum, based on phpBB and customized at CIDR, used for documenting equipment maintenance and service calls. Each new lab instrument is entered into the CIDR Troubleshooter upon arrival. Any subsequent service on or problem with the piece of equipment is added as a new topic for that instrument. Service center and service technician contact information, serial number, and laboratory location of the piece of equipment are stored as well.

**Exemplar LIMS:** Our implementation of the Sapio Sciences Exemplar LIMS, customized extensively in-house, supports all NGS workflow (library construction, target capture, and sequencing) information tracking. This third-party LIMS solution allows for rapid creation of user-specific data types and incorporation of these elements into tasks and workflows modeled after validated wet-

bench protocols developed for production lab use. These workflows create sequential tasks that record information about reagent lot numbers, robotics used, technician, and task completion timestamp. Each task can be made to enforce predefined order to ensure that tasks are not skipped or performed out of order in the lab. Exemplar LIMS contains native validations and reporting elements but also has a robust Java API that allows JHG developers to create custom plugins to extend validations and tasks.